DEVELOPMENT OF A WEB-ENABLED INFORMATICS PLATFORM FOR MANIPULATION OF GENE EXPRESSION DATA

Sheila A Peel[1*], MAJ Karen Kopydlowski[2], and Roland Carel[3], Division of Retrovirology[1], Division of Experimental Therapeutics[2], Walter Reed Army Institute of Research[1,2], and 3rd Millennium, Inc[3]

1.0 ABSTRACT

High-throughput gene expression technologies such as Affymetrix and De Novo 2-Color Microarrays generate large scale measurements that require sophisticated bioinformatic platforms for data archival, management, integration, and analysis if researchers are to derive biologically meaningful relationships from the data. The Array Repository Data Analysis System (ARDAS 1.0) at Walter Reed Army Institute of Research is a web-enabled bioinformatic platform consisting of a Laboratory Information Management System (LIMS), an Analysis Information Management System (AIMS), and a Data Warehouse which coordinates activities of the system modules that allows queries across data within the system. The system, developed following successful Phase 1 and 2 SBIRs in conjunction with our commercial partner, 3rd Millennium, Inc, provides the foundation for an innovative enterprise-wide bioinformatics solution for US Army Medical Research and Materiel Command investigators seeking to leverage array technologies to develop vaccines and drugs against diseases and bio-threat agents of military importance.

2.0 IT COMPONENTS

ARDAS 1.0 is designed and configured to allow access over the network from a zero-footprint client. Internet explorer on a client machine connects to an application server through the network. The application server is a Sun Fire 280 R with 6 GB memory and 2 SPARC III CPUs. The application server hosts the Java application as well as the analytical engine. The Oracle database for the system is hosted on a dedicated Sun Fire 280 R with 3 GB memory and 1 SPARC III CPU. Both Sun Fire servers run the operating system Solaris. They have mirrored internal disks (RAID 1) for fault tolerance. The storage for the database is provided by two 436 GB arrays, each with 12 36 GB SCSI disks. The arrays are configured as RAID5 for fault tolerance. The reporting component of ARDAS 1.0 relies on a Windows 2000 server.

2.0 SYSTEM ARCHITECTURE

The system is based on a best practices four-tiered approach that leverages several technologies as shown in Figure 1. The client tier of the system relies on the Dynamic HTML (DHTML) capability of Internet Explorer 6. The application for the system in the middle tier is built using Java 1.4. The application is hosted on a distribution of JBoss and Tomcat combined. Tomcat is used as a servlet container and JBoss provides JMS capability. The user interface of ARDAS 1.0 is built on the Struts framework from the Apache Jakarta project. The analytical engine uses the R language and BioConductor. R is an open source language and environment for statistical computing and graphics. BioConductor is an open source project comprised of a series of R scripts specifically designed to provide analytical processing methods for genomic and expression data. All BioConductor functions are executed from R. Rserve is TCP/IP server which allows other programs to use facilities in R. Rserve is used to mediate the communication between the ARDAS Java application and R. The back end tier is based on an Oracle 9i database. The database is the primary container for the data in the system and provides the data integrity and data management capabilities.

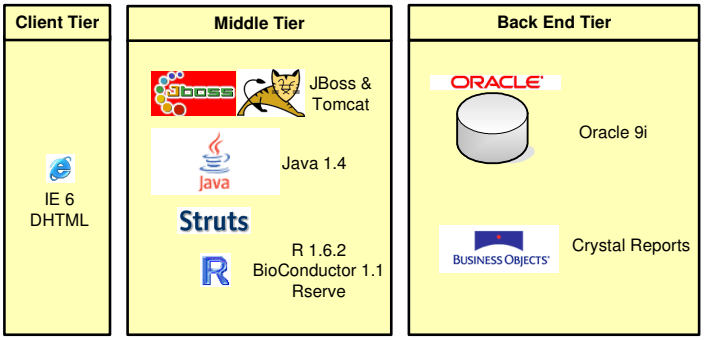

Figure 1. Technologies employed by ARDAS 1.0.

3.0 SYSTEM APPLICATON

The ARDAS application is also built using a best practices tiered approach as shown in Figure 2. With this approach, each tier is responsible for fulfilling a specialized need such as the generation of the user interface or the interaction with the database. The tiered approach separates concerns and makes the system maintainable and extensible. The web-tier in the application is responsible for generating the user interface for the system. The client (Internet Explorer)

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
00 DEC 2004	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Development Of A Web-Enabled Informatics Platform For Manipulation Of Gene Expression Data	**g**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Division of Retrovirology1, Division of Experimental Therapeutics, Walter Reed Army Institute of Research and 3rd Millennium, Inc**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM001736, Proceedings for the Army Science Conference (24th) Held on 29 November - 2 December 2005 in Orlando, Florida., The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**7**	

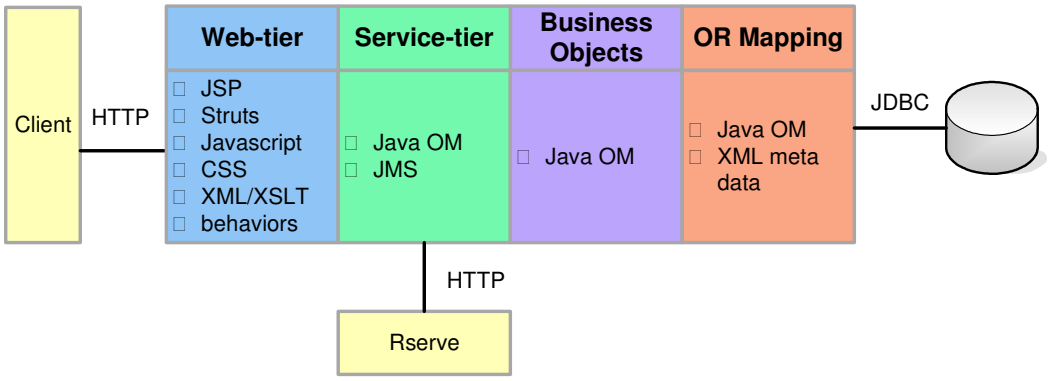

Figure 2. ADRAS 1.0 Application

communicates to the web tier through the application server with HTTP. The user interface is based on DHTML which is itself a combination of several technologies. The pages are built using Java Server Pages and Struts, including the tiles and the validation frameworks. Javascript provides dynamic behavior in the user interfaces. Javascript is also used in DHTML behaviors. User interface components such as tables and trees in ARDAS leverage DHTML behaviors. Cascading style sheets are used extensively to ensure look and feel consistency (e.g., fonts, colors, or sizes) in the user interface. XML/XSLT is used by DHTML components such as the menu bar or the toolbar.

The service tier is concerned with providing complete business functions. For example, the loading of a GPR file involves loading into the file repository, parsing the file into atomic data values, and associating samples to the file. These functions are provided as one service by the business tier. In general, an action in the user interface (e.g., pushing a button or selecting a menu item) invokes an associated service in the service tier. The service tier is based on a Java object model. It also leverages the Java Message Service (JMS) for communicating with Rserve.

The business objects represent fundamental entities and their relationships. For example, ARDAS 1.0 contains business objects for experiment designs, array designs, or biological transcripts. The business objects are implemented in Java.

The Object-Relational (OR) mapping tier converts Java objects into SQL statements and vice-versa. It is responsible for automated conversions between data in Java objects and rows in relational tables. The OR mapping layer is implemented in Java and use XML meta data. The OR layer communicates with the database using JDBC.

4.0 ARDAS 1.O FUNCTIONALITY

ARDAS 1.0 consists of 3 components tightly integrated together in one system: a LIMS, a data Warehouse and search engine, and an AIMS, or Analysis Information Management System.

The LIMS is built around laboratory support for workflows for the production of 2-color spotted data. As one might expect, the result of performing all LIMS associated “tasks” is a detailed audit trail of the series of steps involved in the production of data.

The AIMS includes workflows for normalization and the linear modeling of 2-color spotted data. In the present system configuration linear modeling of Affymetric data is not supported. This feature will be deployed in late 2004.

The Warehouse interfaces with the LIMS and the AIMS, organizes Affymetrix and spotted data, and includes a sophisticated security and data sharing model (see Figure 3).

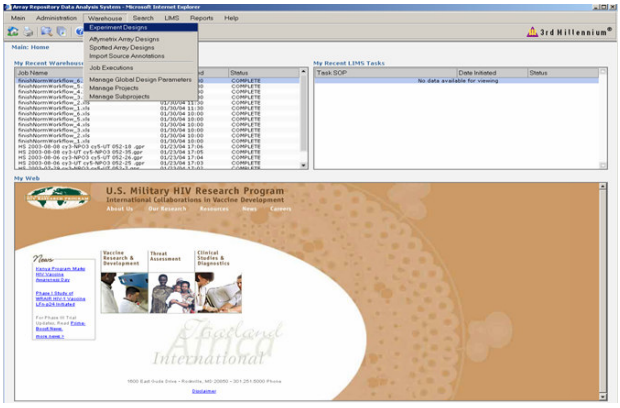

Figure 3. Warehouse screen shot. The Warehouse is the core of ARDAS, and contains all the information derived throughout the experimental and analytical workflows.

The information in the Warehouse is organized into projects which have subprojects containing experimental

designs (see Figure 4). Access privileges are based on projects with 5 levels of security which enables multiple groups at distant locations to collaborate on common projects while protecting information that should not be shared.

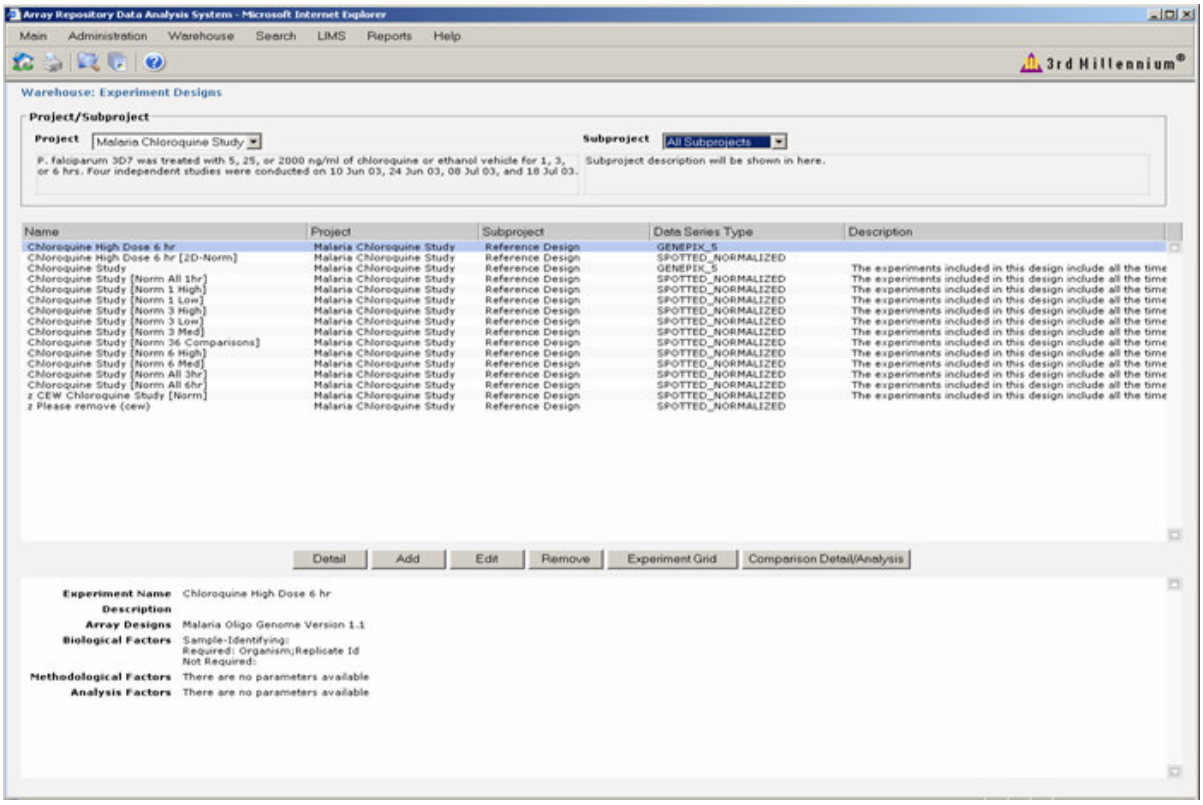

Figure 4. User interface of Experimental Design Project Menu.

An Experimental Design organizes all the expression data collected from microarrays on either Affymetrix one-channel, or two-channel spotted comparisons between samples with given factors (see Figure 5.). Factors include: (1) Biological Factors, the biology of the Samples used to generate the expression data, e.g., organisms or tissue type; (2) Methodological Factors, the physical methods applied to the Samples to generate the data, e.g., time, dose, or extraction protocol; and, (3) Analytical Factors, the analytical methods used to generate the expression data, e.g., normalization or standardization methods. The result is a persistent record within the database of the relationships among parameters (factors), samples, arrays and reporters to which one may add or edit information (see Figure 6). Of note is the capacity for any of this information to become a parameter in the analysis of data.

The AIMS implements analytical workflows based on the aforementioned R and Bioconductor projects. The AIMS currently supports normalization, standardization and linear modeling of two-color spotted array data. An analytical workflow starts in the Warehouse with the selection of a collection of two-color spotted arrays. For each step in a workflow, the user specifies parameters relevant to the step and launches the execution of the functions for the step. This is an iterative process that allows steps to be repeated as many times as necessary to ensure proper analysis of the data. Following completion of the workflow, the data is saved or published in the Warehouse (see Figure 7).

The user interface for the AIMS shows the hierarchy of steps executed upon the data set. All parameters, step results, and the lineage of steps are stored persistently in the database and can be accessed from the Warehouse. Parameters, diagnostic plots, data matrices and summaries are available for each step and generated from open source code. The resultant lists of genes and their expression values can then be stored back into the Warehouse and queried to build biological models of gene expression (see Figure 8).

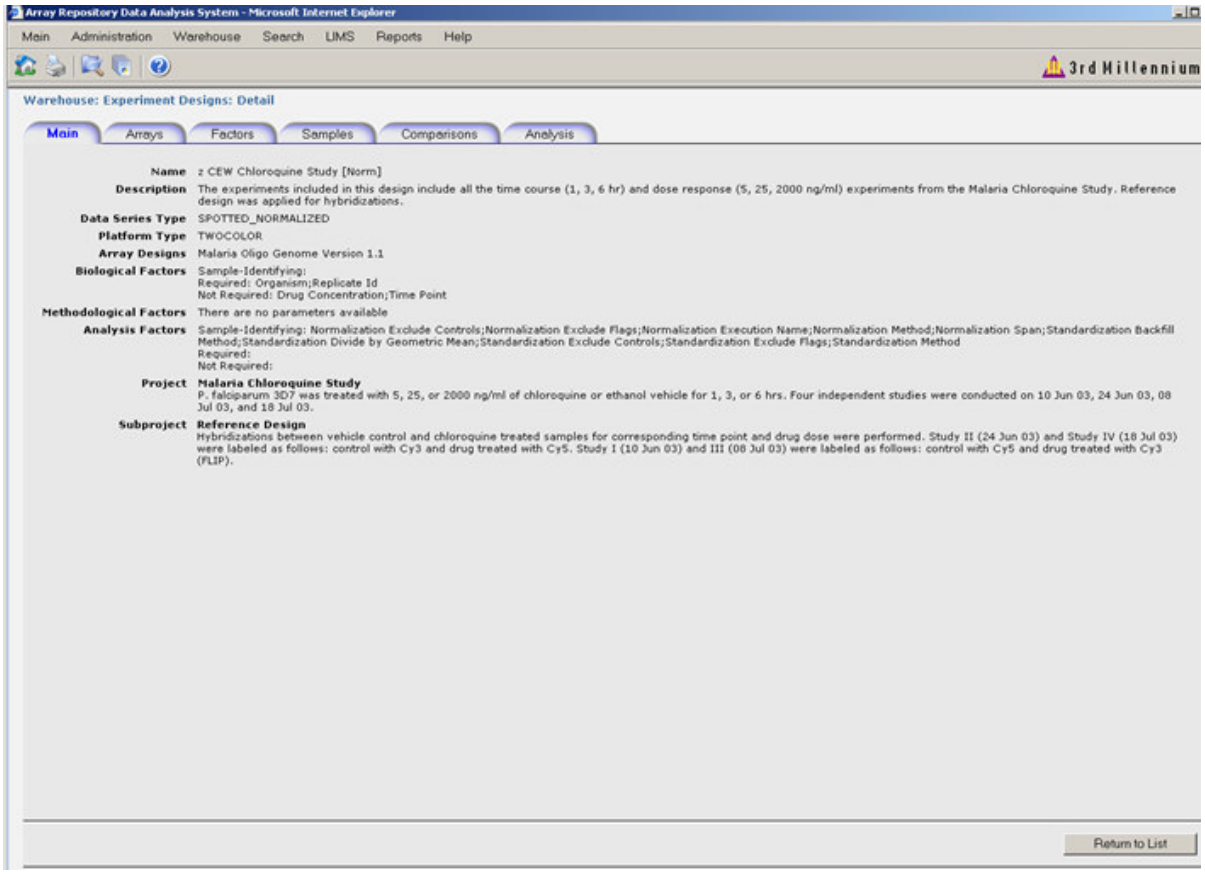

Figure 5. User interface of Experimental Design Detail.

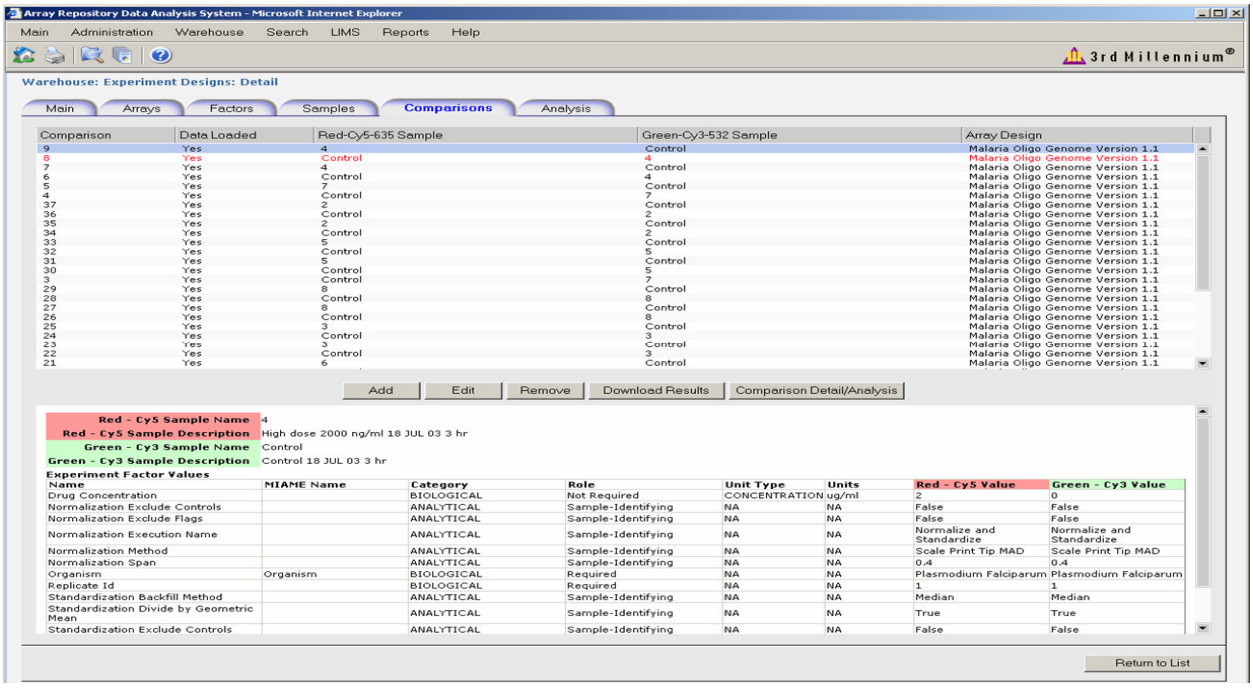

Figure 6. User interface of Experimental Design section of the Warehouse showing a detailed record of the relationships among parameters (factors), samples, arrays and reporters.

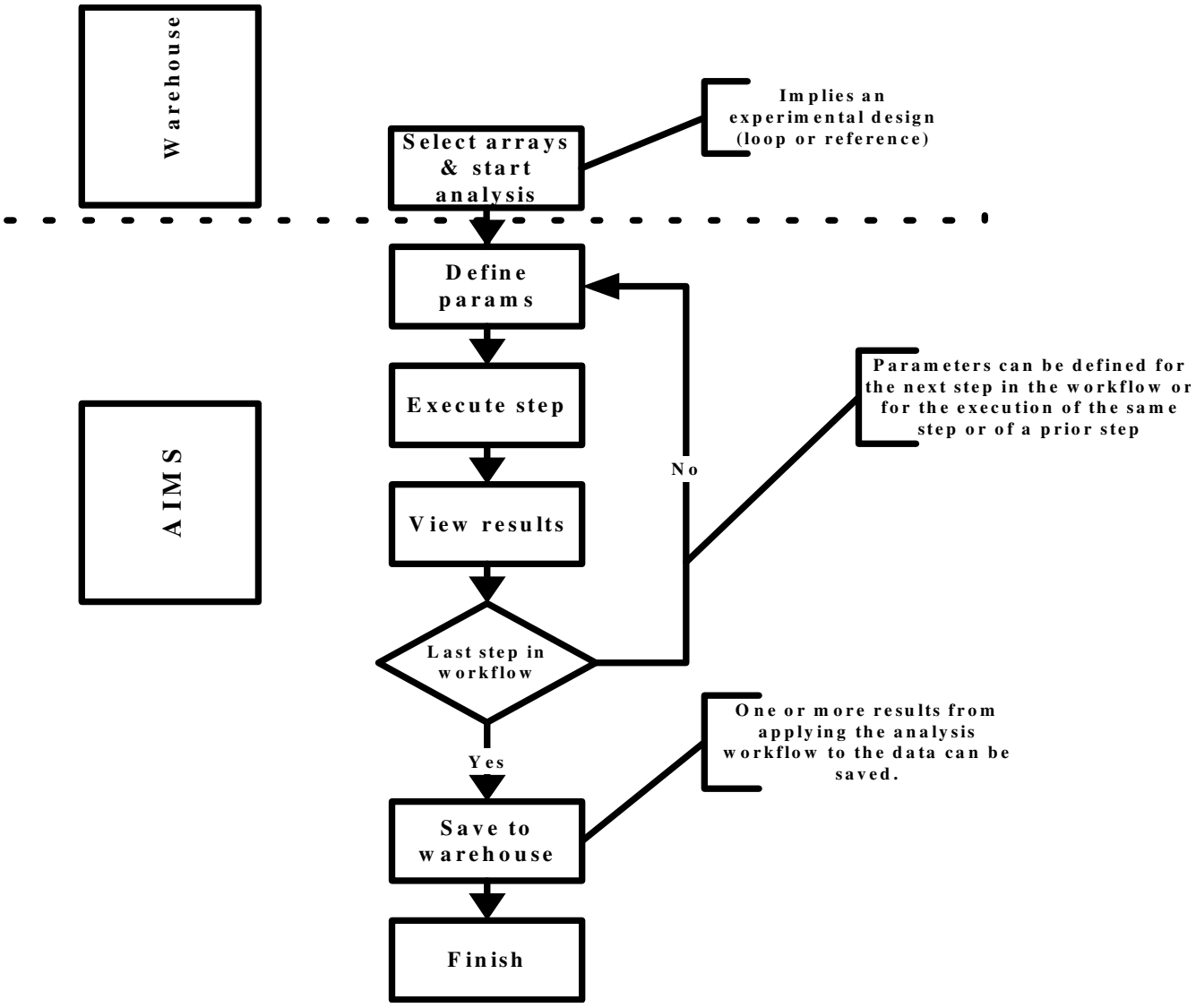

Figure 7. The AIMS, or Analytical Information Management System, implements analytical workflows that guide users through a series of steps which may be performed in an iterative manner.

One of the most powerful features of ARDAS 1.0 is the search engine which uses the relationships among factors, samples, arrays, and AIMS results to form complex queries of both Spotted or Affymetrix data (see Figures 9). One can conduct meta-analyses across experiments and search either independently or simultaneously on gene expression profiles, biological context, and/or gene annotation or functional group (see Figure 10). The results of these complex queries can then be saved in the Warehouse and used in standard set operations to determine either the union, intersection, disjoint, or difference in expression profiles.
The results of the set operations can be saved to the Warehouse as custom user sets for future analyses (see Figure 11).

Lastly, scientific journals now require that authors reporting results of gene expression experiments submit their microarray data in a standardized format, defined as MIAME (Minimal Information About a Microarray Experiment), to public data repositories. Not only is ARDAS 1.0 fully MIAME compliant in that all aspects of gene expression experiments captured and compiled according to the standardized MIAME checklist, the system exports the data electronically to the public repository, Array Express, at European Bioinformatics Institute in Cambridge, UK (see Figure 12).

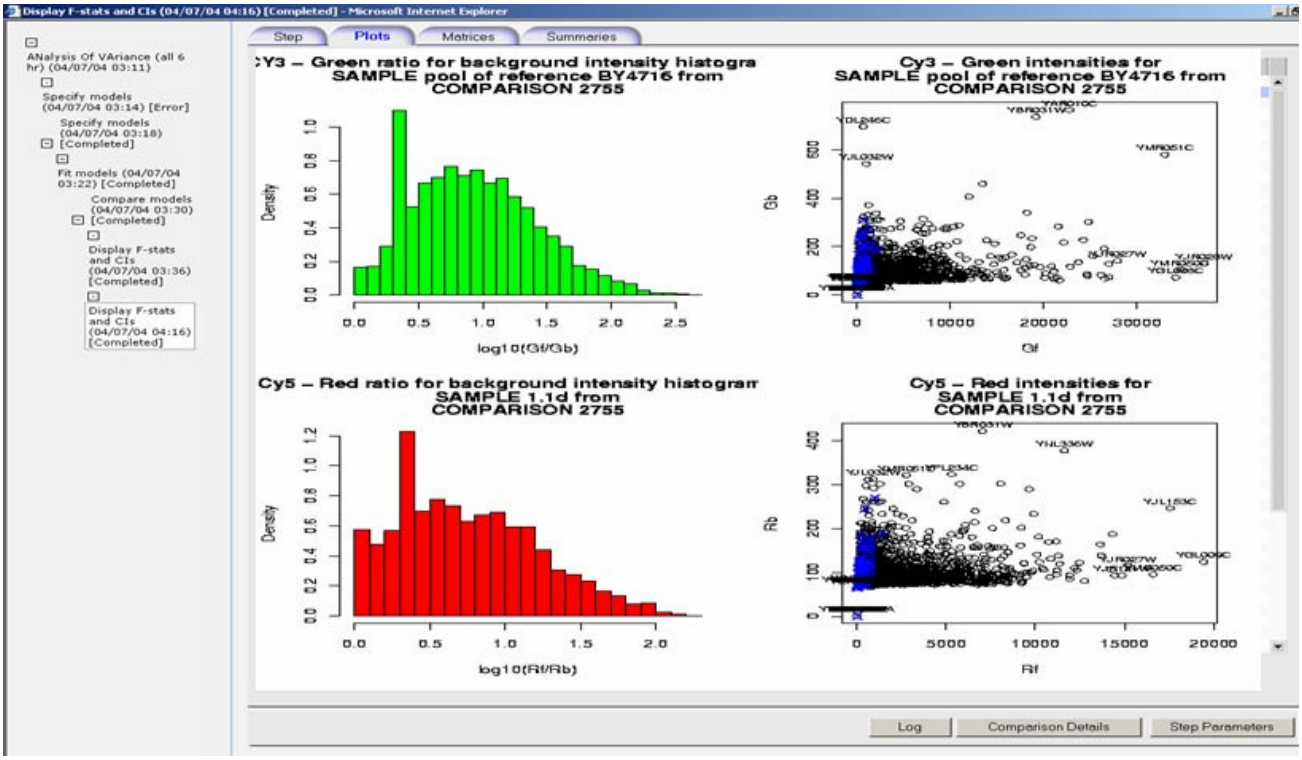

Figure 8. The user interface for the AIMS shows the hierarchy of steps executed upon the data set. Plots are generated by open source code.

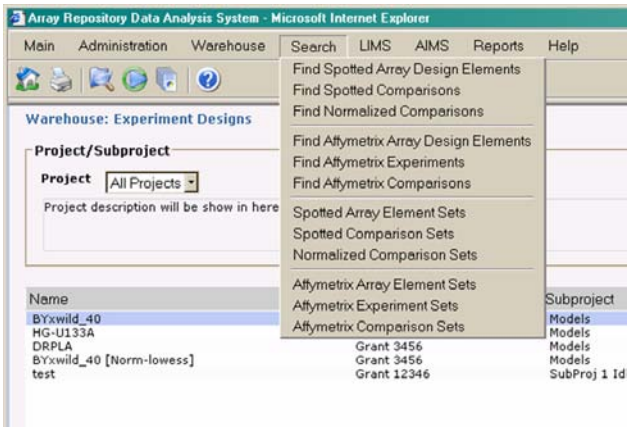

Figure 9. User interface for Search feature of ARDAS 1.0.

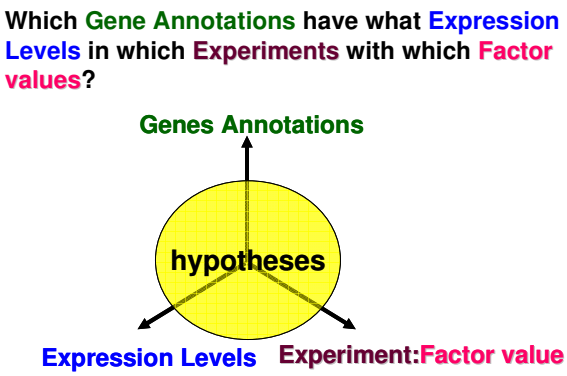

Figure 10. Complex queries can be performed using gene expression values, experimental factors, and or gene annotations.

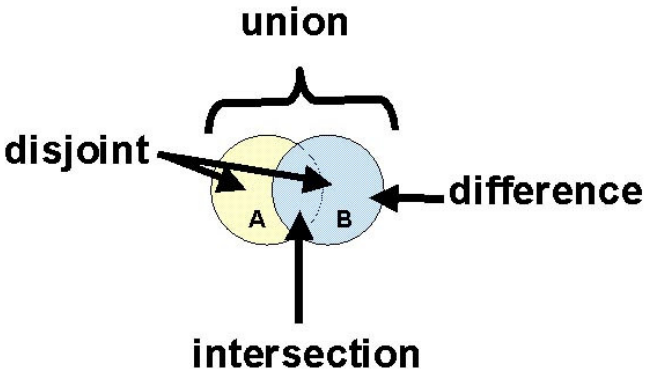

Figure 11. The results of complex queries can be used in standard set operations, then published to the Warehouse to be used in subsequent analyses.

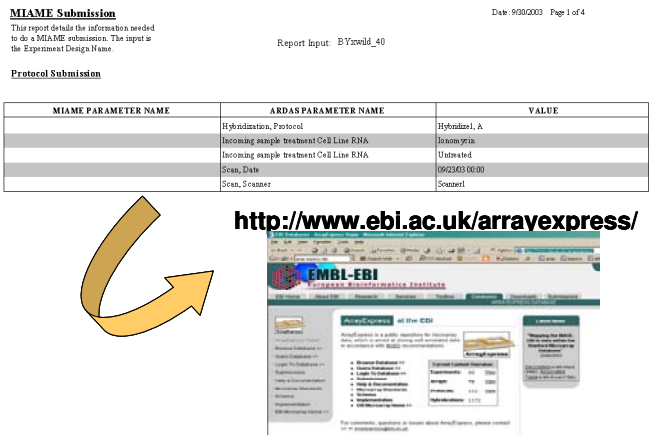

Figure 12. User interface for compiling reports, and electronic MIAME submissions to Array Express.

5.0 SUMMARY

ARDAS 1.0 is a web-enabled bioinformatic platform based on the enterprise technologies, Oracle, Java, and JMS. The platform leverages open source code for data acquisition, data management, and data analysis from Affymetrix and 2-color spotted array platforms. The systems' extensible, modular component-based approach allows system expansion to other functional genomic platforms such as metabolomics and proteomics, and to federated databases for knowledge management.

A successful SBIR Phase I completed during second quarter 2004 resulted in well defined user requirements for customized expansion of ARDAS 1.0 to allow DoD investigators to rapidly manipulate gene expression data through multiple simultaneous analytical, visualization, annotation, and modeling workflows. The proposed expansion, ARDAS 2.0, will readily assist DoD researchers in building complex gene association models, and expedite time to biological discovery. The system advancements continue to provide critical audit trails through data streams to ensure acceptance of gene expression data in FDA licensure applications.

ACKNOWLEDGMENTS

Material has been reviewed by the Walter Reed Army Institute of Research. There is no objection to its presentation and/or publication. The opinions or assertions contained herein are the private views of the author, and are not to be construed as official, or as reflecting true views of the Department of the Army or the Department of Defense.